VTT Technical Research Centre of Finland

Research themes in big data analytics for policymaking

Suominen, Arho; Hajikhani, Arash

Published in:
Policy and Internet

DOI:
10.1002/poi3.258

Published: 01/12/2021

Document Version
Publisher's final version

Received: 26 October 2020 | Accepted: 15 April 2021

DOI: 10.1002/poi3.258

RESEARCH ARTICLE

P&I Policy & Internet PSO WILEY

Research themes in big data analytics for policymaking: Insights from a mixed-methods systematic literature review

Arho Suominen[1,2] | **Arash Hajikhani**[1]

[1]Quantitative Science and Technology Studies, VTT Technical Research Centre of Finland, Espoo, Finland

[2]Department of Industrial Engineering and Management, Tampere University, Tampere, Finland

Correspondence

Arho Suominen, Quantitative Science and Technology Studies, VTT Technical Research Centre of Finland, Tekniikantie 21, 02044 Espoo, Finland.
Email: arho.suominen@vtt.fi

Funding information

H2020 Society, Grant/Award Number: 870822

Abstract

The use of big data and data analytics are slowly emerging in public policy-making, and there are calls for systematic reviews and research agendas focusing on the impacts that big data and analytics have on policy processes. This paper examines the nascent field of big data and data analytics in public policy by reviewing the literature with bibliometric and qualitative analyses. The study encompassed scientific publications gathered from SCOPUS ($N = 538$). Nine bibliographically coupled clusters were identified, with the three largest clusters being big data's impact on the policy cycle, data-based decision-making, and productivity. Through the qualitative coding of the literature, our study highlights the core of the discussions and proposes a research agenda for further studies.

KEYWORDS

bibliometric, big data, content analysis, data analytics, public policy

INTRODUCTION

Big data and data analytics have been seen as augmenting knowledge, ultimately leading to better decision-making. Arguments such as that the broad-based use of big data and data analytics will lead to the end-of-theory speak volumes about our expectations of big data and data analytics technologies' transformative power. While industry has been leading the way to test big data and analytics, public actors have been slower to engage (Poel et al., 2018), despite an equal opportunity for big data and data analytics to augment the public policy process.

Utilizing big data and data analytics has become a near necessity due to our increasing capability for creating and collecting data at an extraordinary rate. The terms "big data" and "data analytics" have been among the buzzwords of recent years, leading to an upsurge in research, industry, and government applications (Zhou et al., 2014). The increased interest in big data in public policy can be seen in the scientific literature Figure 1, highlighting the increase in big data and analytics related literature. We also see public organizations increasingly engaging with big data analytics to solve challenges like the sustainability crisis and pandemics.[1] Scholarly discourse has highlighted case studies and narratives on implementing big data and data analytics in the policy process. However, the literature lacks a systematic view of the current state of big data and data analytics in public policy, and there are identifiable research gaps (Desouza & Jacob, 2017).

This article presents findings from a study of government policy, big data, and data analytics. Using a mixed-methods approach, we analyzed a data set of 538 recent articles to uncover clusters of research agendas focusing on different aspects of big data and data analytic use in the policy cycle. The study's objective is to offer insights into the public policy aspects of analytics, big data, and decision-making. While technological tools are central, this study focuses on the human-centric aspects of big data and data analytics. The study offers a view of the foundations of big data and data analytics in the public policy literature, enabling scholars to have a more substantial and holistic viewpoint. We focused on two research questions:

RQ1. What are the thematic communities of big data and data analytics literature concerning public policy-making?

RQ2. What are the research questions emerging under each of the thematic research communities?

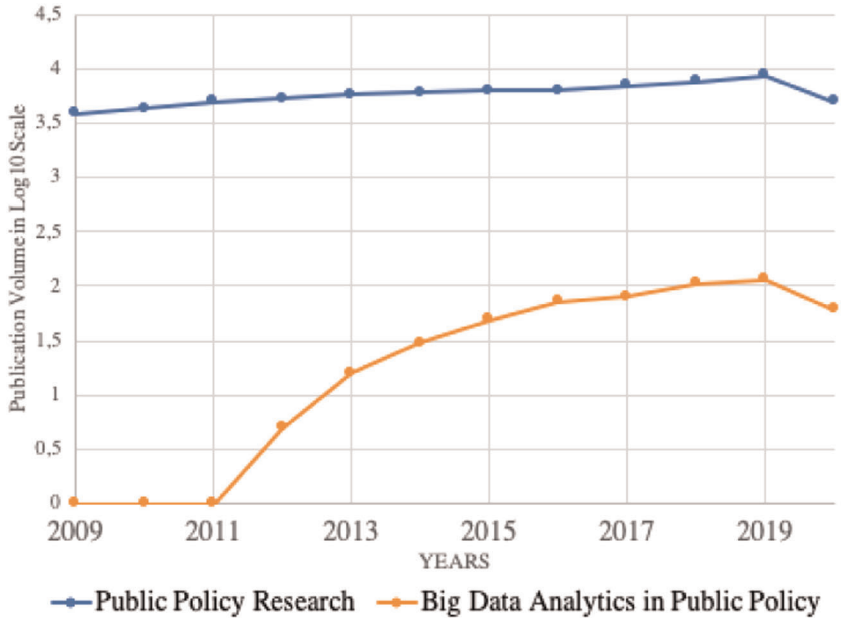

FIGURE 1 Comparison of big data and data analytics focused literature compared to the overall literature. Note that the publications volume is shown in log scale

Our study adopted a mixed-method systematic literature review approach based on a robust empirical bibliometric analysis followed by a qualitative analysis of the core documents to answer these questions. Using a well-established bibliometric method, bibliographic coupling, we identified thematic differences within the literature, and here, we highlight points of departure from the extant literature. The bibliometric analysis was, in turn, used as a basis for the qualitative analysis of the core literature, which is used to propose a research agenda.

We find nine contemporary research communities addressing different aspects of big data and data analytics in public policy. While these communities have significant overlap, our analysis identifies them drawing from different theoretical foundations. Moreover, we demonstrate three larger research strands taking different vantage points, namely building strategic capability, data-based decision-making and productivity increases. Finally, our work proposes a research agenda focusing on the role of strategic capability, data-based decision-making, how to address expectations for better services while simultaneously increasing productivity and how to leverage policy analytics and empiricism.

Our results offer scholars in public policy a vantage point to the theoretical foundations of research in big data and data analytics in public policy-making. We also draw from the identified communities to highlight emerging research themes that can guide research forward. For policymakers, our results highlight the on-going scholarly debate that focuses on addressing critical issues in the adoption of big data in public policy-making, namely capability building and the extent of data-based decision-making. This article will proceed as follows: next, we review the central elements of big data in policy-making. This is followed by a description of the data and our mixed-method approach. Finally, the empirical results are described and followed by a discussion to make sense of the research themes emerging from the analysis.

BACKGROUND

"Big data" is a general term used for the process of gathering massive amounts of data from different sources. Sources can include human-input data but also includes data from sensors or different types of monitoring systems that create process data while running. It is clear that we are accumulating data at a never before seen rate. Already, in 2014, the pace was staggering, with 90% of the world's data being collected during the prior 2 years and 2.5 quintillion bytes of data added each day (Kim et al., 2014). Having access to massive amounts of data has enabled significant innovation in both the public and private domains. Looking at companies like Google and Amazon, with their innovation of new services for consumers, or at the recent ability for doctors to detect cancer cells more precisely thanks to massive training data about what a cancerous cell is, we can see that we are very much on the cusp of creating a broad utility of big data and analytics. This has been seen as a shift in the Industrial Revolution's magnitude (Richards & King, 2014) and has been widely hyped in business (Margetts & Sutcliffe, 2013). That said, public policy is not at the forefront of the use of big data and data analytics in decision making (Kaski et al., 2019; Poel et al., 2018). This nonadoption is due to multiple factors limiting these technologies' utility (Malomo & Sena, 2017).

The ever-increasing amount of data offers possibilities for discovering new relationships and inferencing a multitude of problems. However, this comes with new challenges involving reproducibility, complexity, security, and risks to privacy and a need for new technology and human skills. This is very much the case in public policy, where we need to clearly identify where big data can add value in an ethical and trustworthy manner. In a review, Giest (2017) highlighted three underlying factors to consider. First, institutional capacities have a

significant role in the use of big data in public policy, producing solutions that can enable users to easily interact with data while also taking into account the siloed data structures in the public domain. However, we know from previous research that siloed structures are an important limiting factor for public policy utilization of big data (Malomo & Sena, 2017). Second, hand-in-hand with big data comes the broader digitalization of public services. Digitalization allows for mediums to interact with big data but also enables the creation of new data. There is, however, evidence that digitalization changes the interactions between citizens and public officials and requires new skills from both parties. Third, big data information will have an impact on the policy cycle. Studies have found that there has been limited progress in taking advantage of big data and analytics (Poel et al., 2018) because it requires a significant change in the policy cycle (Höchtl et al., 2016).

Giest (2017) highlights two issues, the substantive role and the procedural role of big data in policy instruments. Procedural activities focus on regulatory activities, such as enabling open data, while substantive actions relate to collecting data for enhancing, for example, evidence-based policy making. Capacities, digitalization, and the role of big data in the (substantive and procedural) policy cycle are core to digital-era governance and evidence-based policy making. In this, it is important to note that policy-makers are not a homogeneous group, and policy cycles vary. Thus, the objectives of analytics throughout the policy cycle vary significantly (Daniell et al., 2016) whether or not we approach the policy cycle as separate discrete stages (Jann & Wegrich, 2007), and it has been shown that big data analytics, when used more in some policy stages than in others, notably improved government transparency, policy evaluation, foresight, and agenda setting (Poel et al., 2018). This should be reflected against findings that data analytics have been politically significant in all policy cycle stages (Van der Voort et al., 2019).

To overcome the challenges, Poel et al. (2015) highlighted multiple topics that must be addressed to enable capacity building, digitalization, and data integration into the policy cycle. These are (1) a skills gap, (2) reduced transparency due to data analytics, (3) sources and tools, (4) standardization of methods and tools, (5) linking of policy experiments with impact assessments, and (6) enabling policy-makers to be informed about the tools that are developed and piloted. The highlighted themes give context to the issue of big data in policy. While we see the significant impacts being created by the use of big data in policy making, along with the subsequent adaptation of data analytics, we need to better explain and make transparent the utility and complementarity of big data-driven analyses for the policy cycle (Vydra & Klievink, 2019).

The challenge highlighted by Poel et al. (2015) and Giest et al. (2017), is also reflected in Pencheva et al. (2020) and Ingram (2019). Both note that big data in public policy has focused more on the "techno-rational factors," dismissing the importance of interaction with the policy process. We know that technology adoption is dependent on the perceived usefulness by the user (Venkatesh & Davis, 2000) and that there is scepticism towards the use of big data and data analytics is public policy-making (Guenduez et al., 2020). This can be the result of a mismatch in practise and expectations. Durrant et al. (2018) show how there is a aspirational motivation to the use of big data and data analytics, not reflected by its everyday utility. While we know that by employing data drive approaches, there is significant potential for anticipatory governance, interaction among stakeholders is the key to draw value from big data and analytics (Maffei et al., 2020; Starke & Lünich, 2020). While the increased stakeholder involvement does not protect from big data and data analytics policy-making creating hard to detect inequalities (Giest & Samuels, 2020; van Veenstra et al., in press). However, engaging with a large pool of stakeholders in the public policy-making process will increase the complexities of adopting big data and data analytics (Janssen et al., 2017).

In addition to stakeholder interaction, the ability to build capacity and also evaluate deficiencies is important (Okuyucu & Yavuz, 2020). Building capacity should not be merely seen as the technical capacity, while that is also important (Poel et al., 2018), but a more holistic capability to integrate big data and analytics into the policy cycle (Höchtl et al., 2016). Policy-making organizations, while being exceptionally technically capable, can be in a situation where the benefits from big data and data analytics remain small due to "applications" not fitting "their organizations and main statutory tasks" (Klievink et al., 2017). This to say that when we talk about big data and data analytics capabilities in public policy-making, the literature focuses on technical issues but also on the ability of big data and data analytics to produce policy-relevant applications.

While we see an increasing and diverse set of research addressing different challenges of big data and data analytics, the current body of literature lacks holistic research agendas (Desouza & Jacob, 2017) addressing the issues highlighted from practice by Giest (2017) and Poel et al. (2015). While we can note emerging fields such as policy analytics (De Marchi et al., 2016; Tsoukias et al., 2013), there is a need to better understand the theoretical grounding and research gap of big data and data analytics in public policy-making.

METHODS AND DATA

This study's methodological approach was based on a mixed method of quantitative and qualitative analyses of the bibliometric data and the publication's content. The selected four-step mixed-methods approach, described below, enables a holistic approach to comprehending the current state-of-the-art and allows us to propose an agenda for going forward. The first phase focused on retrieving the sample of relevant articles and their bibliometric data for analysis. The second phase involved the bibliometric analysis of the retrieved data, which was performed by analysing descriptive statistics, bibliographical coupling, network analytics, and community detection. By gaining a comprehensive view of the more extensive body of literature, we could implement more filtering process based on eigenvector centrality to get a shortlist of papers for the next phase. The third phase, qualitative analysis, continued the process with an in-depth review and coding of the articles' full text. Finally, in the fourth phase, Synthesis, we draw insights from the MAXQDA coding analysis and reporting. This four-phase process is shown in Figure 2.

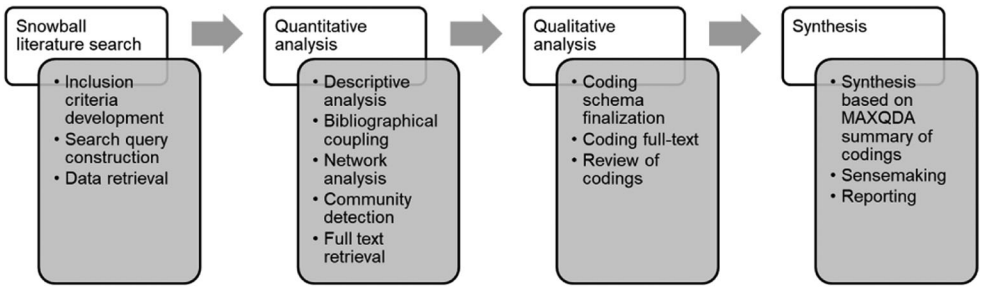

FIGURE 2 The four-step literature review methodological process

Identification of the relevant literature

The data used in this study were retrieved from the Scopus database. Scopus is Elsevier's abstract and citation database that has over 1.7 billion cited references dating back to 1970. A central aspect of the quality of the results is that the query used to search for relevant articles was correctly designed. The study focuses on public policy-making and big data and data analytics. The study's scope is relatively narrow, focusing solely on publications that address policy-making and the policy process. The decision of the scope excludes articled that focus on, for example, big data or data analytics, but lack the specific aspect of policy-making. This was the key inclusion criteria of articles into the data set.

To focus on this specific scope, we used an iterative approach where multiple search strings were tested, and after each search, the abstracts of the 10 most cited articles and the 10 most recent articles were reviewed to understand if the query results reflected the objectives of the study. In practice, the process started with a seed query of "big data" or "data analytics" and "public policy." The query results were reviewed to estimate which articles focused on big data or data analytics and policy-making. These articles were reviewed to see if new terms emerged through the titles, abstracts, and keywords that needed to be included in the analysis. The process adjusted based on a subjective evaluation of the number of false-positives in the 10 most recent and 10 most cited publications and the number of articles retrieved. This method of short-listing the important literature is known as the snowball method, and the process includes consulting the bibliographies of the key documents to find other relevant titles in the subject (Jalali & Wohlin, 2012). After multiple tests of a comprehensive query that also limited the number of false-positives, we downloaded the metadata for 538 documents. These documents were retrieved using the query "public policy," "policy analysis," "policy making," or "public administration," with the terms "big data," "data analytics," or "automated decision-making" in the title, abstract, or keywords of the document.

Quantitative analysis using bibliometrics

To analyze the literature, we used the well-established bibliometric method of bibliographical coupling. Bibliographical coupling allows for analysis of the publications' shared intellectual background (Kessler, 1963), highlighting contemporary research (Youtie et al., 2013). It is an approach to analyzing the shared theoretical background of scientific publications where the link between documents is calculated by the number of references the two documents share. Kessler (1963) elaborates, "A single item of reference shared by two documents is defined as a unit of coupling between them," and if multiple references are shared, the weight of the coupling increases. Bibliographical coupling is able to highlight hot topics (Glanzel & Czerwon, 1996) and links documents with a similar research focus (Jarneving, 2007), ultimately creating a "contemporaneous representation of knowledge" (Youtie et al., 2013). This approach has been used in several research papers to form the basis for research agenda building (Suominen et al., 2019; Yuan et al., 2015).

Using the retrieved publication metadata, the VOSviewer tool (van Eck & Waltman, 2009) was selected to calculate bibliographical coupling weights for all the documents in our data set. VOSviewer is a free tool used for bibliometrics and was selected due to the exports available in the software allowing for deeper network analysis in Gephi. The SCOPUS data export was used as an input to the VOSviewer. During the analysis process, we selected documents as the level of analysis, minimum number of citations for a document was set to zero and the full set was selected for the analysis. The full counting method, which assigns each researcher with full credit of one publication rather than a fractional share per the

number of authors, was used for the calculation method. Finally, we accepted VOSviewer default to keep the most extensive set of related items, which limited the analysis to the largest, by node, subgraph created by the bibliographical coupling analysis. This limited the analysis to 332 documents.

Bibliographical coupling analysis of a data set creates a graph $G = (V, E)$, formed by V, a set of nodes, and E, a set of edges joining nodes. As we calculated the link weight e between each publication node v, we created a simple unidirectional graph. This graph data, produced with the VOSviewer tool, was imported to Gephi because it allows for more detailed visualization, network measure calculation, and community detection.

Further network analysis, including network descriptive values, for example, degree, for graph G were calculated in Gephi. Communities were identified using Blondel's (Blondel et al., 2008) fast unfolding networks algorithm. The fast unfolding networks algorithm is one of the most computationally efficient methods to find high modularity partitions of networks in short time. The methods starts by assigning each node to a separate community there after calculating the gain in modularity by merging neighboring nodes in a community. This process is continued through the network, ultimately creating a new network with nodes assigned to communities (see Blondel et al., 2008, for a detailed explanation). In the Gephi software, the modularity algorithm can be controlled by a resolution variable that controls the number of communities the algorithm creates. This variable was changed to limit the number of tiny clusters. We increased the resolution value until even the smallest community has approximately one percent share of the documents.

We also calculated the eigenvector centrality for each document in the data set. In graph theory, eigenvector centrality measures the influence a node v has in G. A value is calculated to all v based on an idea that connections to other important nodes are more important than equal and/or low-scoring nodes. This centrality value describes a publication's relevance to the overall network created by the bibliographical coupling analysis. Filtering for communities, we identified the five most eigenvalue central publications from each community to be selected for systematic coding. Selecting the five most central publications was done to keep the final sample of documents in the coding phase manageable, and selecting the most eigenvector central publications was a method to take the most important publications from each community for a deeper analysis.

Qualitative analysis of the literature

Central to creating meaningful implications through content analysis is to create valid coding schemes. The created schema is key to generating confidence in the results of the content analysis. Due to our mixed-method approach, the content analysis was already based on a rigorous bibliometric method that identified research themes. The content analysis's central element was to uncover latent features in the community, using the five most eigenvector central documents in each community that could shed light on the created cluster's theme. While there is no right way to do content analysis (Weber, 1990), the coding protocol is central to reproducible results. To this end, Gaur and Kumar (2018) offered a four-quadrant framework by topical area, being either a method or a research theme and the scope being narrow or wide. The current study is focused on a narrow field and focuses on research themes drawn from the bibliometric data. For this type of content analysis, Gaur and Kumar (2018) proposed a seven-point coding approach, which includes codes for the following:

1. Research subthemes
2. Primary variables
3. Scope of study

4. Context of study
5. Conceptual or empirical study
6. Theory(ies) employed
7. Key findings

The authors offer the framework as a template to be customized to the study's objective at hand. For the current study, we adopted the majority of elements in the framework creating a eight-point approach including (1) the context of the study, (2) primary variables, and (3) the scope defined as a research gap. We merged the conceptual or empirical study and theories employed to be the (4) code employed for the theoretical or methodological framework. We also included a separate code for (5) the method used to understand the specific approach used in the study. For the key findings, we also created coding. The first focused specifically on any concrete (6) results mentioned, and the second focused on essential (7) discussion and (8) conclusions. This was explicitly done to highlight other study results and to inform the content analysis in the scholarly debate around the research themes.

To add additional rigor to the coding process, the qualitative analysis was conducted using MAXQDA software. The software allows for coding documents, directly annotating the documents with the codes, and thereafter drawing syntheses from the created codings. The use of the software increases transparency and the trustworthiness of the analysis (Costa et al., 2017; Sinkovics & Alfoldi, 2012). After the coding schema was created, the practical coding process was completed by one researcher, who read the five eigenvector central full-text documents from each cluster and annotated the documents based on the framework. The second researcher had the role of validating the annotation results. The interaction between the researchers was to makes sure that all of the items in the coding schema were identified. The second researcher went through the papers and codings made by the first researcher to make sure that each schema item, if available, was identified. However, as publications can repeat the same information, our approach was designed to ensure we do not capture the same information multiple times per publication. This selection of an approach made using, for example, intercoding agreement impractical.

The MAXQDA analysis software used in the coding automatically created cross-tabulations of the coded documents and created a synthesis document about the coded sections of text. These were used in the interpretation phase. The synthesis of coded sections provided by MAXQDA was used to interpret the results. The two researcher familiarized with the MAXQDA summaries of the communities independently. After independent review of the synthesis, the researchers discussed on their findings. There after the authors worked jointly to draw insights from the coding results, working toward synthesis of the core areas of future research.

RESULTS

The retrieved publications are recent, with the first publication in the data set published in 2009, as seen in Figure 3. The data were retrieved in early June 2020; thus, publications for 2020 represent the first 5 months, and one should expect a growing trend in publication volume. While Figure 3 shows growth, it is important to put this into context. Figure 1 projects the frequency of the topics of big data and data analytics in the public policy related literature concerning the overall public policy literature for the same duration. The publication volumes are normalized on log 10 scale to be able to illustrate them in one view. It is clear that the body of literature focused on big data and data analytics in public policy is

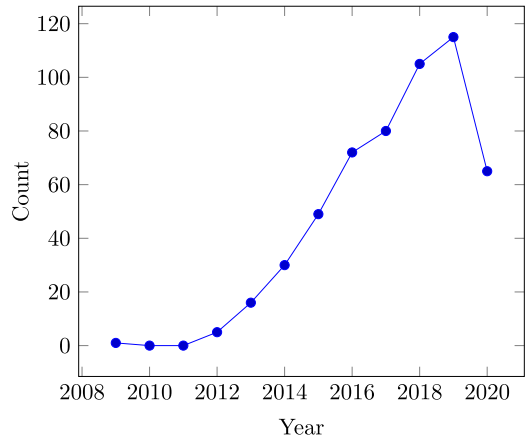

FIGURE 3 Count of publications on a yearly basis. Data for 2020 only reflects the first 5 months

TABLE 1 Count of publications within disciplines

Discipline	Count
Computer Science	280
Social Sciences	196
Engineering	118
Decision Sciences	77
Business, Management, and Accounting	75
Mathematics	60
Medicine	52
Environmental Science	39
Economics, Econometrics, and Finance	27
Energy	25

experiencing a much sharper increase in interest in comparison to the overall public policy body of literature.

Over half of the publications are from computer science, social sciences, or engineering. As seen in Table 1, the three mentioned disciplinary areas have over 100 publications. Table 1 highlights all disciplinary areas with over 20 publications in the data set. Notably, the smaller areas are case-study driven, highlighting the use of big data and data analytics in, for example, the topics of health care and environmental issues and energy.

The descriptive analyses of the data also highlight the different journals that are attracting manuscripts on the topic. As shown in Table 2, the major publication sources for the articles included in the data set are mostly computer and information science journals. Analysing Table 2, we should note our scope. The study focused solely on big data or data analytics and its use in public policy-making. In the listed journals, the number of articles focusing on any single area, for example, big data is much higher. We should also note that the search in SCOPUS only looks at the title, abstract and keywords, making articles discussing big data or data analytics and policy-making in full-text only undiscovered by the

TABLE 2 Count of publications by publication source

Source	Count
ACM International Conference Proceeding Series	34
Lecture Notes in Computer Science[a]	15
Communications in Computer and Information Science	10
Advances in Intelligent Systems and Computing	9
Policy and Internet	9
Journal of Public Health Policy	8
Government Information Quarterly	5
Journal of Cleaner Production	5
Journal of Policy Analysis and Management	5
Public Administration Review	5

[a]Including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics.

TABLE 3 Count of publications by institution

Institution	Count
Delft University of Technology	15
The University of Hong Kong	7
University at Albany	5
The University of Texas at Austin	5
University of Washington, Seattle	5
University of the Aegean	5
Leiden University	5
New York University	5
University of California, Berkeley	5
University of Oxford	5

query. It is notable that the publication sources with at least five publications have, all together, 105 publications, or approximately 19.5% of all publications. This implies that publications are scattered over many different publication sources, and an on-going debate on the subject is hard to pin down to a specific outlet.

The division of the publications by country aligns with that of global scientific publication production, except China, which, with 60 publications, is lagging significantly behind the United States, with 141 publications.[2] The two largest science producers are followed by the United Kingdom (55 publications), Italy (37 publications), Netherlands (29 publications), and India and South Korea (both with 28 publications). Looking at the affiliations of the publications, they are, again, sporadically divided between different organizations. As seen in Table 3, only Delft University of Technology has a significant track record of publications in the data set, and already, starting with the third-highest publication count, there were only five publications per institution.

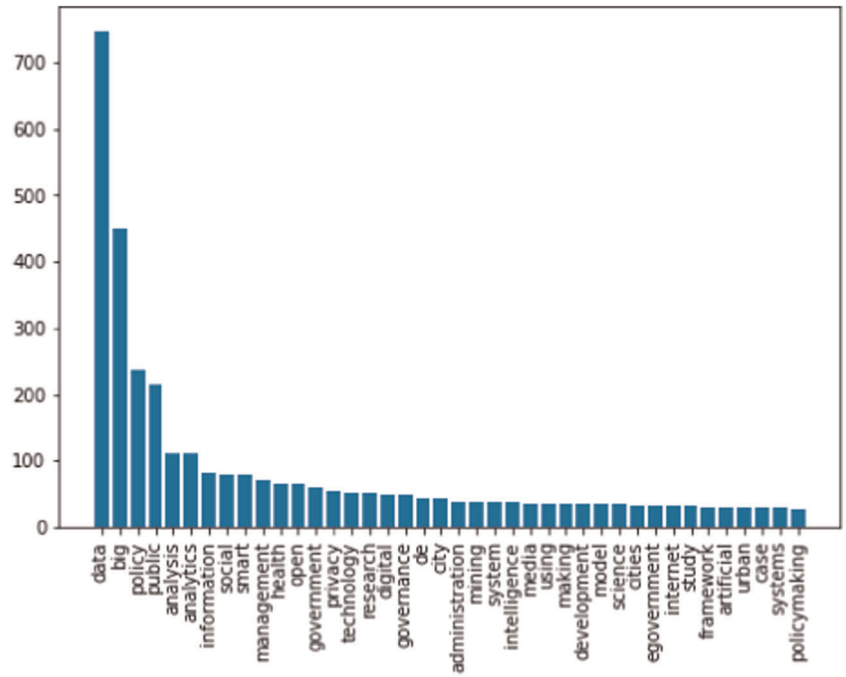

FIGURE 4 Most common terms, truncated at 40 most common, from the title and keyword fields

To understand more in-depth the sample, we concatenated the title and keywords of the publications in the sample. We then cleaned the concatenated title and keyword fields by removing punctuations and English stopwords.[3] Seen in the Figure 4, the terms extracted are as expect, focusing on big data and public policy. More thematic terms emerging to the most frequent terms relate to management, health, cities, and media, giving insight into the articles' context. Overall, the descriptive analyses highlight that the data set contains only recent publications, divided among multiple sources and institutions, with relatively low volumes produced by each. The terms in the publications aligned with the search query.

The 538 publications' bibliometric data were analyzed for bibliographic coupling using VOSviewer software. During the calculation process, the software first analyzed if any of the publications were detached and isolated from the overall network emerging from the data. VOSviewer calculates the largest subnetwork in the full networks and offers an option to only used the largest connected set. Selecting to use the largest connected set allows focusing on the core documents, removing the isolates. In our data set, the largest set of connected items creating a network was 332 documents. The outlier documents were dropped from the final sample, as these articles would not be included in the bibliometric coupling based clusters but would, rather, remain isolates throughout the analysis. Continuing with the cohesive sample of 332 articles, the VOSviewer created network was imported to Gephi software for further analysis. Network metrics were calculated, and the bibliographic coupling network had a weighted degree 22.56, with the density of the network being 0.048. This to say that there is significant linkages between node (documents) in the communities as weighted degree is significant, but the full network is not strongly linked, as the density is low. Next, the communities were detected using the modularity algorithm (Blondel et al., 2008). The modularity algorithm resolution was increased from its default value of one, until the smallest community had an approximately 1% share of the documents. With a resolution

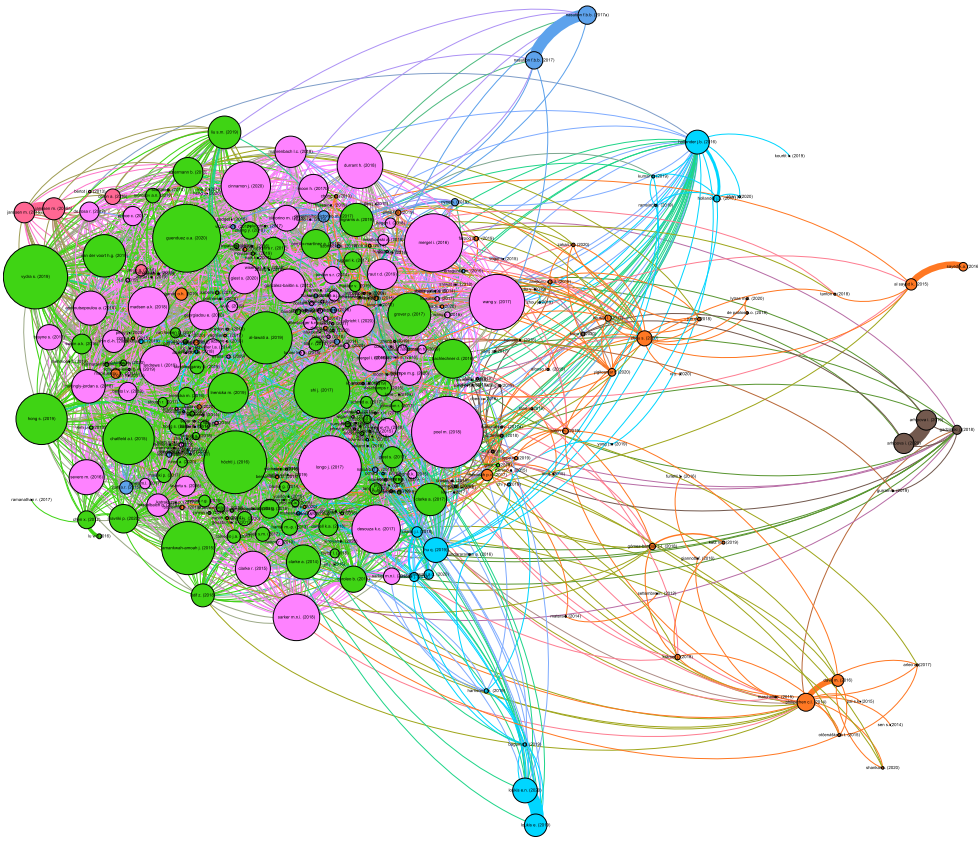

FIGURE 5 Bibliographical coupling network where color highlights the cluster size and citation count of publication

of 1.4, the algorithm resulted in nine communities, with the the smallest community has only 0.9% of the publication. The largest community had 32.23% of the publications.

A visual representation of the created graph can be seen in Figure 5.[4] In this figure, the color shows the clusters created using the modularity algorithm. The size of a node reflects the number of citations of an article in the network. As the graph highlights, the network was essentially created by two large communities visualized in the graph in green and purple. These communities contain 32% and 24% of the documents. In addition, there is a third large cluster, shown in orange, with roughly 20%, but is not visible as a concentrated and highly cited community's of research. In addition to these three larger communities, the network includes smaller communities attached to the center of the network, all with less than 8% of total documents. From the communities, we selected the most eigenvector central documents for further qualitative analysis. For each community, we selected the five most central documents. However, for two smallest communities this resulted in all documents being selected. For the third smallest community, due to the lack of access to documents, we were only able to include four documents. The documents selected can be seen in Table 5. The documents were then individually read and coded using the MAXQDA software.

Seen in Table 4 is the summary of the qualitative analysis of the literature selected. In the Table 4, rows describe the number of coded sections per each of the six-point coding elements. The columns describe the communities #1 being the largest #2 second largest

TABLE 4 Communities by number of codings per area according to MAXQDA software analysis

#	#1	#2	#3	#4	#5	#6	#7	#8	#9	SUM
Context	4	4	5	5	4	3	4	4	2	35
Research gap	4	4	2	3	4	4	4	3	3	31
Framework	3	1	0	2	0	0	0	2	0	8
Variables	0	1	0	0	0	1	0	2	2	6
Method	2	3	1	4	1	2	3	3	2	21
Result	4	3	1	2	1	0	3	3	1	18
Discussion	4	4	1	3	3	1	3	2	1	22
Conclusions	4	4	1	4	3	2	3	2	0	23
Sum	25	24	11	23	16	13	20	21	11	164

and #9 the smallest community. Based on the analysis, Communities 3, 5, 6, and 9 were the least informative for the analysis. We can also note that the articles mostly contributed to understanding the context and research gap. Table 4 highlights that the majority of the articles are narratives and not based on an empirical setting, framework, or variables. This is in line with the reading, where the majority of articles were narratives or reviews of cases. We should note that the Table 4 does not explain the value or informativeness of a community, but highlights the amount of information available for the analysis.

Based on the coding summaries in the communities of research, we identified different vantage points and research gaps. Our analysis highlighted two large highly cited communities of research. The largest community, seen in Figure 5 in green, focused on integrating big data into the policy cycle. Referring to this community as "Big data impacting policy cycle," our analysis identified that publication in this community focused on if the value was captured from the abundance of data create by public actors. While public actors have been amassing and profiting from big data (Wang et al., 2015), these articles put significant emphasis on on technical capability and then taking advantage of the analytics (Höchtl et al., 2016; Vydra & Klievink, 2019). In the articles, the core context is created by the notion that public organizations are looking for ways to take advantage of big data (Chatfield et al., 2015; Guenduez et al., 2020; Vydra & Klievink, 2019). Vydra and Klievink (2019) discussed techno-optimism and policy-pessimism, the notion that big data will provide tools for better decision making, but we might only be selecting the easy-to-handle aspects of big data. In this largest community of research the publications focused on the lack of empirical research (Chatfield et al., 2015), particularly focusing on the technical capacity (Höchtl et al., 2016; Wang et al., 2015) and the lack of actual implementations of big data in public policy (Guenduez et al., 2020; Vydra & Klievink, 2019). This community clearly identified the role of strategic capability (Chatfield et al., 2015), enabling a more dynamic policy cycle (Höchtl et al., 2016) and broad-based knowledge diffusion (Wang et al., 2015), ultimately translating big data from better evidence to better policy (Guenduez et al., 2020; Vydra & Klievink, 2019).

The second largest community, labeled as "Data based decision-making" and seen in Figure 5 in purple, focused particularly on how data informs decision making (Desouza & Jacob, 2017) in more general terms or in specific areas such as education (Wang, 2017), departing from the notion that more data will ultimately make more adaptive (Longo et al., 2017) and evidence-based (Poel et al., 2018) policy possible. This was even to the extent of moving to a posttheoretical phase (Mergel et al., 2016). However, the papers highlight the

TABLE 5 Eigenvector central publications selected for further analysis

%	Label	Titles
32.2	Big data impacting policy cycle	Techno-optimism and Policy-Pessimism in the Public Sector Big Data Debate (Vydra & Klievink, 2019); Technological Frames in Public Administration: What Do Public Managers Think of Big Data? (Guenduez et al., 2020); Safety or No Safety in Numbers? (Wang et al., 2015); Governments, Big Data and Public Policy Formulation; Big Data in the Policy Cycle: Policy Decision Making in the Digital Era (Höchtl et al., 2016); Capability Challenges in Transforming Government through Open and Big Data: Tales of Two Cities (Chatfield et al., 2015)
24.4	Data based decision-making	Big Data in Public Affairs; Big Data for Policymaking: Great Expectations, but with Limited Progress? (Poel et al., 2018); Technology Use, Exposure to Natural Hazards, and Being Digitally Invisible: Implications for Policy Analytics (Longo et al., 2017); Big Data in the Public Sector: Lessons for Practitioners and Scholars (Desouza & Jacob, 2017); Education Policy Research in the Big Data Era: Methodological Frontiers, Misconceptions, and Challenges (Wang, 2017)
19.3	Productivity	To Do More, Better, Faster and More Cheaply: Using Big Data in Public Administration (Maciejewski, 2017); Toward a Political Economy of Nudge: Smart City Variations (Gandy & Nemorin, 2019); Exploring Development of Smart City Research through Perspectives of Governance and Information Systems: A Scientometric Analysis Using CiteSpace (Zhou et al., 2020); Applied Spatial Modelling in the Twenty-First Century: The Wilson Legacy. Looking Back and Looking Forward (Birkin & Clarke, 2019); Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data (Chen & Zhang, 2014)
8.1	Policy analytics	Twitter Data in Public Administration: A Review of Recent Scholarship (Hu, 2019); Social Media in Aid of Post Disaster Management (Malawani et al., 2020); Artificial Intelligence-Based Public Sector Data Analytics for Economic Crisis Policymaking (Loukis et al., 2020); Negotiating the Reuse of Health-Data: Research, Big Data, and the European General Data Protection Regulation (Starkbaum & Felt, 2019); Change Calls Upon Public Administrators to Act, But in What Way? Exploring Administration as a Platform for Governance (Zingale et al., 2018)
6.0	IoT and public policy	Urban Big Data and Sustainable Development Goals: Challenges and Opportunities (Kharrazi et al., 2016); Mobile Phone Data Statistics as a Dynamic Proxy Indicator in Assessing Regional Economic Activity and Human Commuting Patterns (Arhipova et al., 2020); Birth of Industry 5.0: Making Sense of Big Data with Artificial Intelligence, "the Internet of Things" and Next-Generation Technology Policy (Özdemir & Hekim, 2018); Digital Government, Smart Cities and Sustainable Development (Zheng et al., 2019); Perspectives of the Use of Smartphones in Travel Behaviour Studies: Findings from a Literature Review and a Pilot Study (Gadziński, 2018)
5.7	Value of data	Government Information Policy in the Era of Big Data (Washington, 2014); Big Data: What Is It and What Does It Mean for Cardiovascular Research and Prevention Policy (Pah et al., 2015); Conceptual Framework for Public Policymaking Based on System Dynamics and Big Data (Nasution & Bazin, 2017); Adoption of Green Electricity Policies: Investigating the Role of Environmental Attitudes via Big

TABLE 5 (Continued)

%	Label	Titles
		Data-Driven Search-Queries (Lee et al., 2016); Big Data Applications in Health Sciences and Epidemiology (Pyne et al., 2015)
2.1	E-Government	Big Data and Public Policy: Can It Succeed Where e-Participation Has Failed? (Bright & Margetts, 2016); The "Social Side" of Public Policy: Monitoring Online Public Opinion and Its Mobilization During the Policy Cycle (Ceron & Negri, 2016); The Evolution of Information and Communication Technology in Public Administration (Liu & Yuan, 2015); Big Data and e-Government: Issues, Policies, and Recommendations (Bertot & Choi, 2013)
1.2	Impact asssessment	Governance by Targets and the Performance of Cross-Sector Partnerships: Do Partner Diversity and Partnership Capabilities Matter? (Alonso & Andrews, 2019); Policy Analytics and Accountability Mechanisms: Judging the 'Value for Money' of Policy Implementation (Scharaschkin & McBride, 2016); Bridging Big Data and Policy Making: A Case Study of Failure (Kudo, 2018); Reputation as Public Policy for Internet Security: A Field Study (Tang et al., 2012)
0.9	Implementation	Incorporation of Social Media Indicator in e-Government Index(Wahid et al., 2019); Value of Telecom Operators' Big Data in Social Public Management (Hong et al., 2020); A New Dimension in Urban Planning: The Big Data as a Source for Shared Indicators of Discomfort (Scattoni et al., 2014)

Note: "%" refers to share of documents included to the community from the all documents.

lack of concrete implementation of big data (Desouza & Jacob, 2017; Wang, 2017). With the nascent implementation, the studies also highlight the negative impact on invisible sub-populations (Longo et al., 2017), inaccuracies in gathered data (Poel et al., 2018), and the overall hurdles of gathering, retaining, and analyzing data (Mergel et al., 2016). There is a clear need for a systematic research agenda (Desouza & Jacob, 2017) to fully grasp the potential of big data, and this agenda should consider the transparency of data (Poel et al., 2018) and inclusiveness (Longo et al., 2017; Mergel et al., 2016).

The above mentioned two communities, seen in the Figure 5 in green and purple, include over 50% of the articles and are the highest cited as seen from the size of the nodes. Community 3 focused on "Productivity" increases in public management (Chen & Zhang, 2014), enabling better service for the public (Maciejewski, 2017). The articles focused particularly on the context of urban surroundings, smart cities (Gandy & Nemorin, 2019; Zhou et al., 2020), and transportation (Birkin & Clarke, 2019). This community had more narrative focusing on particular cases, but the bibliometric review by Zhou et al. (2020) highlighted a disconnect between information systems research and governance research.

Community 4 focused on new forms of "Policy analytics" (Loukis et al., 2020), relying on new forms of empiricism (Starkbaum & Felt, 2019) and collaborative action (Zingale et al., 2018), such as through social media (Hu, 2019; Malawani et al., 2020). This community identified a gap in the strategic use of available data and the role of citizens engaging unfiltered via social media (Zingale et al., 2018) or data donations in health care (Starkbaum & Felt, 2019). For the novel data sets, these works highlight the importance of creating facilitating conditions (Loukis et al., 2020; Malawani et al., 2020) and deeper understanding of the benefits (Hu, 2019; Starkbaum & Felt, 2019) and impacts of data sets, particularly social media.

Community 5, labeled "IoT and Public policy," focused particularly on the Internet of Things and smart cities and highlighted the need for governance frameworks (Kharrazi

et al., 2016; Özdemir & Hekim, 2018) and policies (Arhipova et al., 2020; Zheng et al., 2019), as well as the need to invest in the data infrastructure (Kharrazi et al., 2016). Community 6 focused on the "Value of data" in public policy. Nasution and Bazin (2017) highlighted the importance of gathering data and creating links between the different data sources available to public actors (Pah et al., 2015), but also highlighted the implications of secondary use of public data (Washington, 2014). This community of papers focused on the health care sector, highlighting the potential impact but also the failures of big data and policy analytics (Pah et al., 2015). But with the large potential (Pah et al., 2015) comes a need to form information policy to make use of the data (Washington, 2014).

Community 7 focused on "E-government" and the transformation enabled by digital government services (Bertot & Choi, 2013). Highlighting the increase in e-government (Liu & Yuan, 2015), particularly its power in transparency and interaction (Ceron & Negri, 2016; Bright & Margetts, 2016; Bertot & Choi, 2013), the research gap in this community of papers focused on the depth of e-government adoption (Liu & Yuan, 2015) and its ability to influence the policy process (Bright & Margetts, 2016). There is need to develop information and communication technologies (ICT) innovations that foster collaboration between citizens and government (Bertot & Choi, 2013), while also supporting the policy process (Liu & Yuan, 2015). Community 8 focused on the "Impact assessment" of security (Tang et al., 2012), public actions (Kudo, 2018), policy analytics, and contractual governance (Alonso & Andrews, 2019). Contributions in this community were narrative but highlighted the need for impact assessment. Community 9, labeled "Implementation," focused on data in urban planning (Hong et al., 2020; Scattoni et al., 2014), with limited linkages to big data and analytics.

DISCUSSION AND CONCLUSION

The present study's main finding highlighted the communities of research appearing to embrace big data and analytics in public policy. The analysis focused on nine contemporary communities of research addressing different aspects of big data and data analytics in public policy. While the communities had significant overlap, the bibliographical coupling analysis showed that they had different theoretical origins. Some of this might be the result of citation biases due to, for example, the practises in citing prolific papers or self-citations, but the results suggest that research on the topic is sparse and lacks a cohesive foundation.

The results show two large communities with a significant number of citations and a third relatively large community with less citations and scattered across the bibliographic coupling network. These three approach big data and data analytics in public policy from different vantage points: building strategic capability, data-based decision making and productivity increases. The literature on productivity is more scattered, which can partly stem from the fact that productivity can be addressed from multiple vantage points, for example, macroeconomic productivity or task productivity. The smaller communities in the analysis remain somewhat detached from the three larger communities, focusing on case study contributions drawn from the broader body of literature. However, Community 4 is engaging as it discusses policy analytics as a clearly defined area of research and application. This suggests an emerging terminology in the field that can draw the scattered literature into a more cohesive grounding.

For the future development of these recently emergent research strands, we suggest a research agenda that builds cohesion among the communities. This study agenda's core questions depart from the three largest communities supported by the smaller areas. One of the most essential areas of development related to building capabilities. Capabilities have

already been identified as an important area when considering using big data and data analytics in public policy-making (Giest, 2017). We should note that when we talk about capabilities, literature looks at capabilities in various ways ranging from strategic to the high operational vantage point. Future research should continue focusing on the role of strategic capability (Chatfield et al., 2015) while not losing sight of the importance of technical capacity (Höchtl et al., 2016; Wang et al., 2015). Future work on capabilities should also be able to use empirical evidence draw insights (Guenduez et al., 2020; Vydra & Klievink, 2019), focusing mainly on the issue about how to reduce the skills gap and enable a balanced approach toward stakeholders influence (Washington, 2014) and engagement (Bright & Margetts, 2016; Bertot & Choi, 2013).

The second avenue of future research should focus on different aspects of data-based decision-making. Within this study stream, the discussion highlights general data-based decision-making issues and uses, for example, artificial intelligence. We know that at a general level, ethics of data-based decision-making has seen significant interest (Herschel & Miori, 2017; Zwitter, 2014), and public policy-making is not an outlier in this. Particular aspects of future research should address transparency (Poel et al., 2018), inclusiveness (Longo et al., 2017; Mergel et al., 2016), and interaction between the relevant stakeholders (Bertot & Choi, 2013; Bright & Margetts, 2016). This is also an area that would be well-served by developing a systematic research agenda (Desouza & Jacob, 2017). A third, while smaller and not so concentrated, area of research focuses on productivity. In this, we see an interplay of two issues. We should better understand how big data and data analytics can address expectations for better service for the public (Maciejewski, 2017) while simultaneously increasing productivity (Chen & Zhang, 2014). The promise of big data in policy-making is that we can provide better service while increasing the productivity of the work. This tension is an important avenue of research.

The fourth and final, avenue of research focuses on policy analytics. One can question if we could reduce the phrase "big data and data analytics in public policymaking" to "policy analytics." This is a justified question and there is literature to support the term "policy analytics" defining the field (Daniell et al., 2016; De Marchi et al., 2016; Tsoukias et al., 2013), but our analysis still positions in as an emerging theme. This said, the community has the possibility to draw the scattered literature into a more cohesive grounding. To make this a reality, research should address new forms of policy analytics (Loukis et al., 2020) and empiricism (Starkbaum & Felt, 2019) by understanding facilitating conditions (Loukis et al., 2020; Malawani et al., 2020), impacts to the policy cycle (Bright & Margetts, 2016; Höchtl et al., 2016), and impact assessment (Scharaschkin & McBride, 2016). This is a broad scope of research, but one that could, under as cohesive approach, create significant impacts.

This study presented a literature review based on bibliometric analysis and qualitative analysis of the core documents. While the mixed method approach is robust, there are two limitations to consider. First, the data set was gathered from one archival source of publications, which, while having good coverage, probably does not capture the comprehensive scholarly literature that comes as reports and not necessarily scientific publications. Also, there is a considerable delay in updating conference publications in the database. Second, bibliographic coupling uses citations, and therefore depends on citation practices. The method does not, for example, consider the strength of a particular citation or the general increase in the number of references in academic literature.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 870822.

ENDNOTES

[1]For example, efforts by the OECD on transforming policy in, for example, COVID http://www.oecd.org/competition/big-data-bringing-competition-policy-to-the-digital-era.htm or United Nations activites on SDG https://www.un.org/en/sections/issues-depth/big-data-sustainable-development/index.html.

[2]China's share of global scientific publishing is 20.67%, the United States' is 16.54%, followed by India, Germany, Japan, United Kingdom, Russia, and Italy, all with a share of less than 6%.

[3]For stopwords we used the NLTK package in Python.

[4]The figure is also available as an interactive network graph at https://determined-montalcini-938eb3.netlify.app/.

ORCID

Arho Suominen http://orcid.org/0000-0001-9844-7799
Arash Hajikhani https://orcid.org/0000-0003-2032-9180

REFERENCES

Alonso, J. M., & Andrews, R. (2019). Governance by targets and the performance of cross-sector partnerships: Do partner diversity and partnership capabilities matter? *Strategic Management Journal*, *40*(4), 556–579.

Arhipova, I., Berzins, G., Brekis, E., Binde, J., Opmanis, M., Erglis, A., & Ansonska, E. (2020). Mobile phone data statistics as a dynamic proxy indicator in assessing regional economic activity and human commuting patterns. *Expert Systems*, *37*, e12530.

Bertot, J. C., & Choi, H. (2013). Big data and e-government: Issues, policies, and recommendations. In *Proceedings of the 14th Annual International Conference On Digital Government Research* (pp. 1–10).

Birkin, M., & Clarke, M. (2019). Applied spatial modelling in the twenty-first century: The wilson legacy. looking back and looking forward. *Interdisciplinary Science Reviews*, *44*(3–4), 286–300.

Blondel, V. V. D., Guillaume, J.-L. J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.

Bright, J., & Margetts, H. (2016). Big data and public policy: Can it succeed where e-participation has failed? *Policy & Internet*, *8*(3), 218–224.

Ceron, A., & Negri, F. (2016). The "social side" of public policy: Monitoring online public opinion and its mobilization during the policy cycle. *Policy & Internet*, *8*(2), 131–147.

Chatfield, A., Reddick, C., & Al-Zubaidi, W. (2015). Capability challenges in transforming government through open and big data: Tales of two cities. In *Proceedings of Thirty Sixth International Conference on Information Systems*. Fort Worth, TX, USA.

Chen, C. P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, *275*, 314–347.

Costa, A. P., de Sousa, F. N., Moreira, A., & de Souza, D. N. (2017). Research through design: Qualitative analysis to evaluate the usability. In *Computer Supported Qualitative Research* (pp. 1–12). Springer.

Daniell, K. A., Morton, A., & Insua, D. R. (2016). Policy analysis and policy analytics. *Annals of Operations Research*, *236*(1), 1–13.

De Marchi, G., Lucertini, G., & Tsoukiàs, A. (2016). From evidence-based policy making to policy analytics. *Annals of Operations Research*, *236*(1), 15–38.

Desouza, K. C., & Jacob, B. (2017). Big data in the public sector: Lessons for practitioners and scholars. *Administration & Society*, *49*(7), 1043–1064.

Durrant, H., Barnett, J., & Rempel, E. S. (2018). Realising the benefits of integrated data for local policymaking: Rhetoric versus reality. *Politics and Governance*, *6*(4), 18–28.

Gadziński, J. (2018). Perspectives of the use of smartphones in travel behaviour studies: Findings from a literature review and a pilot study. *Transportation Research Part C: Emerging Technologies*, *88*, 74–86.

Gandy Jr., O. H., & Nemorin, S. (2019). Toward a political economy of nudge: Smart city variations. *Information, Communication & Society*, *22*(14), 2112–2126.

Gaur, A., & Kumar, M. (2018). A systematic approach to conducting review studies: An assessment of content analysis in 25 years of IB research. *Journal of World Business*, *53*(2), 280–289.

Giest, S. (2017). Big data for policymaking: Fad or fasttrack? *Policy Sciences*, *50*(3), 367–382.

Giest, S., & Samuels, A. (2020). 'For good measure': Data gaps in a big data world. *Policy Sciences*, *53*(3), 559–569.

Glanzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, *37*(2), 195–221.

Guenduez, A. A., Mettler, T., & Schedler, K. (2020). Technological frames in public administration: What do public managers think of big data? *Government Information Quarterly*, *37*(1), 101406.
Herschel, R., & Miori, V. M. (2017). Ethics & big data. *Technology in Society*, *49*, 31–36.
Höchtl, J., Parycek, P., & Schöllhammer, R. (2016). Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce*, *26*(1–2), 147–169.
Hong, Y., Li, Z., & Wang, J. (2020). Value of telecom operators' big data in social public management. In *Journal of Physics: Conference Series* (Vol. 1437, p. 012068). IOP Publishing.
Hu, Q. (2019). Twitter data in public administration: A review of recent scholarship. *International Journal of Organization Theory & Behavior*, *22*(2), 209–221.
Ingrams, A. (2019). Public values in the age of big data: A public information perspective. *Policy & Internet*, *11*(2), 128–148.
Jalali, S., & Wohlin, C. (2012). Systematic literature studies: Database searches vs. backward snowballing. In *Proceedings of the 2012 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement* (pp. 29–38). IEEE.
Jann, W., & Wegrich, K. (2007). Theories of the policy cycle. *Handbook of Public Policy Analysis: Theory, Politics, and Methods*, *125*, 43–62.
Janssen, M., Konopnicki, D., Snowdon, J. L., & Ojo, A. (2017). Driving public sector innovation using big and open linked data (bold). *Information Systems Frontiers*, *19*(2), 189–195.
Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics*, *1*(4), 287–307.
Kaski, S., Ailisto, H., & Suominen, A. (2019). International ai experts: Towards the third wave of artificial intelligence. In *Leading the way into the age of artificial intelligence: Final report of Finland's Artificial Intelligence Programme 2019* (pp. 28–42). Ministry of Economic Affairs and Employment.
Kessler, M. (1963). An experimental study of bibliographic coupling between technical papers (Corresp.). *IEEE Transactions on Information Theory*, *9*(1), 49–51.
Kharrazi, A., Qin, H., & Zhang, Y. (2016). Urban big data and sustainable development goals: Challenges and opportunities. *Sustainability*, *8*(12), 1293.
Kim, G.-H., Trimi, S., & Chung, J.-H. (2014). Big-data applications in the government sector. *Communications of the ACM*, *57*(3), 78–85.
Klievink, B., Romijn, B.-J., Cunningham, S., & de Bruijn, H. (2017). Big data in the public sector: Uncertainties and readiness. *Information Systems Frontiers*, *19*(2), 267–283.
Kudo, H. (2018). Bridging big data and policy making: A case study of failure. In *Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance* (pp. 609–615).
Lee, D., Kim, M., & Lee, J. (2016). Adoption of green electricity policies: Investigating the role of environmental attitudes via big data-driven search-queries. *Energy Policy*, *90*, 187–201.
Liu, S. M., & Yuan, Q. (2015). The evolution of information and communication technology in public administration. *Public Administration and Development*, *35*(2), 140–151.
Longo, J., Kuras, E., Smith, H., Hondula, D. M., & Johnston, E. (2017). Technology use, exposure to natural hazards, and being digitally invisible: Implications for policy analytics. *Policy & Internet*, *9*(1), 76–108.
Loukis, E. N., Maragoudakis, M., & Kyriakou, N. (2020). Artificial intelligence-based public sector data analytics for economic crisis policymaking. *Transforming Government: People, Process and Policy*, *14*(4), 639–662.
Maciejewski, M. (2017). To do more, better, faster and more cheaply: Using big data in public administration. *International Review of Administrative Sciences*, *83*(1_suppl), 120–135.
Maffei, S., Leoni, F., & Villari, B. (2020). Data-driven anticipatory governance. Emerging scenarios in data for policy practices. *Policy Design and Practice*, *3*(2), 123–134.
Malawani, A. D., Nurmandi, A., Purnomo, E. P., & Rahman, T. (2020). Social media in aid of post disaster management. *Transforming Government: People, Process and Policy*.
Malomo, F., & Sena, V. (2017). Data intelligence for local government? assessing the benefits and barriers to use of big data in the public sector. *Policy & Internet*, *9*(1), 7–27.
Margetts, H., & Sutcliffe, D. (2013). Addressing the policy challenges and opportunities of “big data”. *Policy & Internet*, *5*(2), 139–146.
Mergel, I., Rethemeyer, R. K., & Isett, K. (2016). Big data in public affairs. *Public Administration Review*, *76*(6), 928–937.
Nasution, F. B. B., Bazin, N. E. N., & Hasanuddin (2017). Conceptual framework for public policymaking based on system dynamics and big data. In *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)* (pp. 1–7). IEEE.
Okuyucu, A., & Yavuz, N. (2020). Big data maturity models for the public sector: A review of state and organizational level models. *Transforming Government: People, Process and Policy*, *14*(4), 681–699.

Özdemir, V., & Hekim, N. (2018). Birth of industry 5.0: Making sense of big data with artificial intelligence, "the internet of things" and next-generation technology policy. *Omics: A Journal of Integrative Biology*, *22*(1), 65–76.

Pah, A., Rasmussen-Torvik, L., Goel, S., Greenland, P., & Kho, A. (2015). Big data: What is it and what does it mean for cardiovascular research and prevention policy. *Current Cardiovascular Risk Reports*, *9*(1), 424.

Pencheva, I., Esteve, M., & Mikhaylov, S. J. (2020). Big data and AI—A transformational shift for government: So, what next for research? *Public Policy and Administration*, *35*(1), 24–44.

Poel, M., Meyer, E. T., & Schroeder, R. (2018). Big data for policymaking: Great expectations, but with limited progress? *Policy & Internet*, *10*(3), 347–367.

Poel, M., Schroeder, R., Treperman, J., Rubinstein, M., Meyer, E., Mahieu, B., Scholten, C., & Svetachova, M. (2015). *Data for policy: A study of big data and other innovative data-driven approaches for evidence-informed policymaking*. Report about the State-of-the-Art. Amsterdam: Technopolis. Oxford Internet Institute, Center for European Policy Studies.

Pyne, S., Vullikanti, A. K. S., & Marathe, M. V. (2015). Big data applications in health sciences and epidemiology. In *Handbook of statistics* (Vol. 33, pp. 171–202). Elsevier.

Richards, N. M., & King, J. H. (2014). Big data ethics. *Wake Forest Law Review, 49*, 393.

Scattoni, P., Lazzarotti, R., Lombardi, M., Neri, A. R., Turi, R., & Verratti, J. A. Z. (2014). A new dimension in urban planning: The big data as a source for shared indicators of discomfort. *Italian Journal of Planning Practice*, *4*(1), 102–120.

Scharaschkin, A., & McBride, T. (2016). Policy analytics and accountability mechanisms: Judging the 'value for money' of policy implementation. *Annals of Operations Research*, *236*(1), 39–56.

Sinkovics, R. R., & Alfoldi, E. A. (2012). Progressive focusing and trustworthiness in qualitative research. *Management International Review*, *52*(6), 817–845.

Starkbaum, J., & Felt, U. (2019). Negotiating the reuse of health-data: Research, big data, and the european general data protection regulation. *Big Data & Society*, *6*(2), 1–12.

Starke, C., & Lünich, M. (2020). Artificial intelligence for political decision-making in the european union: Effects on citizens' perceptions of input, throughput, and output legitimacy. *Data & Policy*, *2*, e1–e16.

Suominen, A., Seppänen, M., & Dedehayir, O. (2019). A bibliometric review on innovation systems and ecosystems: A research agenda. *European Journal of Innovation Management*, *22*(2), 335–360.

Tang, Q., Linden, L. L., Quarterman, J. S., & Whinston, A. (2012). Reputation as public policy for internet security: A field study. In *Thirty Third International Conference on Information Systems*. Orlando, FL, USA.

Tsoukias, A., Montibeller, G., Lucertini, G., & Belton, V. (2013). Policy analytics: An agenda for research and practice. *EURO Journal on Decision Processes*, *1*(1–2), 115–134.

Van der Voort, H., Klievink, A., Arnaboldi, M., & Meijer, A. J. (2019). Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making? *Government Information Quarterly*, *36*(1), 27–38.

van Eck, N., & Waltman, L. (2009). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.

van Veenstra, A. F., Grommé, F., & Djafari, S. (in press). The use of public sector data analytics in the Netherlands. *Transforming Government: People, Process and Policy*. https://doi.org/10.1108/TG-09-2019-0095

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, *46*(2), 186–204.

Vydra, S., & Klievink, B. (2019). Techno-optimism and policy-pessimism in the public sector big data debate. *Government Information Quarterly*, *36*(4), 101383.

Wahid, J. A., Shi, L., Tao, Y., Wei, L., & Saleem, K. (2019). Incorporation of social media indicator in e-government index. In *Proceedings of the 5th International Conference on Communication and Information Processing* (pp. 201–209).

Wang, X., White, L., Chen, X., & Amankwah-Amoah, J. (2015). Safety or no safety in numbers? Governments, big data and public policy formulation. *Industrial Management & Data Systems*, *115*(9), 1569–1603.

Wang, Y. (2017). Education policy research in the big data era: Methodological frontiers, misconceptions, and challenges. *Education Policy Analysis Archives*, *25*(94), 1–24. https://doi.org/10.14507/epaa.25.3037

Washington, A. L. (2014). Government information policy in the era of big data. *Review of Policy Research*, *31*(4), 319–325.

Weber, R. P. (1990). *Basic content analysis* (Vol. 49). Sage.

Youtie, J., Kay, L., & Melkers, J. (2013). Bibliographic coupling and network analysis to assess knowledge coalescence in a research center environment. *Research Evaluation*, *22*(3), 145–156.

Yuan, Y., Gretzel, U., & Tseng, Y.-H. (2015). Revealing the nature of contemporary tourism research: Extracting common subject areas through bibliographic coupling. *International Journal of Tourism Research*, *17*(5), 417–431.

Zheng, L., Kwok, W.-M., Aquaro, V., & Qi, X. (2019). Digital government, smart cities and sustainable development. In *Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance* (pp. 291–301).

Zhou, S., Zhang, X., Liu, J., Zhang, K., & Zhao, Y. (2020). Exploring development of smart city research through perspectives of governance and information systems: A scientometric analysis using citespace. *Journal of Science and Technology Policy Management*, *11*(4), 431–454.

Zhou, Z.-H., Chawla, N. V., Jin, Y., & Williams, G. J. (2014). Big data opportunities and challenges: Discussions from data analytics perspectives [discussion forum]. *IEEE Computational Intelligence Magazine*, *9*(4), 62–74.

Zingale, N. C., Cook, D., & Mazanec, M. (2018). Change calls upon public administrators to act, but in what way? Exploring administration as a platform for governance. *Administrative Theory & Praxis*, *40*(3), 180–199.

Zwitter, A. (2014). Big data ethics. *Big Data & Society*, *1*(2), 1–6.

How to cite this article: Suominen, A., & Hajikhani, A. (2021). Research themes in big data analytics for policymaking: Insights from a mixed-methods systematic literature review. *Policy & Internet*, 1–21. https://doi.org/10.1002/poi3.258